Une nouvelle méthode de classification monoclasse à base de représentation symbolique : Application à la classification de documents

Nicolas Sidère, Fahimeh Alaei, Nathalie Girard, Sabine Barrat, Jean-Yves Ramel

Une nouvelle méthode de classification monoclasse à base de représentation symbolique : Application à la classification de documents.

Nicolas Sidère[1] Fahimeh Alaei[1] Nathalie Girard[1] Sabine Barrat[1] Jean-Yves Ramel[1]

[1] Université François Rabelais Tours, LI EA 6300, 37200 Tours, France

{nicolas.sidere, fahimeh.alaei, nathalie.girard, sabine.barrat, jean-yves.ramel}@univ-tours.fr

Résumé

Construire un classificateur obtenant de bons résultats tout en utilisant un faible nombre d'exemples d'apprentissage est un besoin récurrent dans le domaine de la classification d'images de documents, et en particulier pour l'entreprise pour laquelle cette étude a été réalisée. Dans ce cas-là, le choix d'utiliser un classificateur monoclasse (nécessitant uniquement des exemples positifs) représente une alternative intéressante. Dans cet article, nous présentons une nouvelle méthode de classification monoclasse basée sur une représentation symbolique. Initialement, un ensemble de caractéristiques est extrait à partir des données de l'ensemble d'apprentissage. Puis, un vecteur d'intervalles issus de ces caractéristiques est construit pour représenter la classe. Chaque intervalle (donnée symbolique) est calculé à partir de la moyenne et de l'écart-type de chaque caractéristique. Pour évaluer le classificateur monoclasse proposé, nous avons utilisé un ensemble de données composé de 544 images de documents. Les expérimentations montrent que le classificateur monoclasse est performant lorsque le nombre d'exemples d'apprentissage est faible (=10). Il est donc utilisable dans un contexte de classification de documents, avec de meilleurs résultats que ceux obtenus par un classificateur k-ppv.

Mots Clef

Classification monoclasse, représentation symbolique, classification d'images de document.

Abstract

Training a system using a small number of instances to obtain accurate recognition/classification is a crucial need in document classification domain, especially for the firm in which we conducted this study. The one-class classification is chosen since only positive samples are available for the training. In this paper, a new one-class classification method based on symbolic representation method is proposed. Initially a set of features is extracted from the training set. A set of intervals valued symbolic feature vector is then used to represent the class. Each interval value (symbolic data) is computed using mean and standard deviation of the corresponding feature values. To evaluate the proposed one-class classification method a dataset composed of 544 document images was used. Experiment results reveal that the proposed one-class classification method works well even when the number of training samples is small (=10). Moreover, we noted that the proposed one-class classification method is suitable for document classification and provides better result compared to one-class k-nearest neighbor (k-NN) classifier.

Keywords

One-class classification, Symbolic data representation, Document image classification.

1 Introduction

Pour ce travail de recherche, nous avons collaboré avec une entreprise française experte en analyse et classification d'images de documents. Un des axes majeurs de leurs travaux porte sur le problème de la classification dans le cas où seulement un faible nombre d'exemples (10 exemples) est disponible pour l'apprentissage d'une classe de document. De plus, dans la majorité des cas, les classes des documents sont définies suivant une information sémantique ne correspondant pas forcément à une logique informatique. Dans un tel contexte, le risque de chevauchement entre classes (au sens informatique) est fort. Par conséquent, la classification s'avère être une tâche complexe. L'utilisation de classificateur monoclasse est une solution pour répondre à cette problématique. Ce choix est d'autant plus intéressant qu'il permet l'introduction de nouvelles classes sans contraindre le système à un nouvel apprentissage même après plusieurs mois de mise en production (évolution continue) ; la maintenance du système s'en trouve alors simplifiée. À partir des contraintes précédemment décrites (faible nombre d'exemples d'apprentissage, risque de chevauchement des classes), nous proposons un nouveau classificateur monoclasse applicable à la classification d'images de documents dans un contexte industriel. L'article est organisé comme suit : dans la section 2, nous exposons un état de l'art sur les classificateurs monoclasse. La section 3 décrit notre méthode basée sur une représentation symbolique. Les expérimentations et comparatifs sont présentés dans la section 4. Nous concluons ensuite l'ar-

ticle et proposons quelques perspectives que nous souhaiterions mener sur ce travail.

2 Classification Monoclasse

En général, l'objectif des algorithmes de classification consiste à classifier un objet inconnu parmi plusieurs classes prédéfinies. Dans le problème de la classification monoclasse, on suppose que seules les données de la classe cible sont disponibles pour l'apprentissage du classificateur, alors que l'ensemble de test comprend des exemples positifs (même classe) et négatifs (classes différentes). Dans la littérature, plusieurs méthodes ont été proposées pour résoudre le problème d'une classification monoclasse [11, 13, 5, 7, 12]. Ces méthodes peuvent être divisées en trois catégories principales [11]. Les méthodes du premier groupe, les méthodes de densité, estiment directement la densité de probabilité des objets positifs [11]. Parmi ces méthodes de représentation, les plus développées sont le modèle gaussien [5] ou le mélange de gaussiennes [8], les modèles de Markov [8], plus proche voisin [11] et estimateurs de densité de Parzen [11]. La deuxième catégorie contient les méthodes à base de frontières dont l'objectif principal est de définir une frontière stable et fiable de la classe [6]. Ces méthodes comprennent la méthode des k-centres [11], la description des données à vecteurs de support (SVDD) [13] et les machines à vecteurs de support (SVM) [6]. Enfin, dans le troisième groupe, les méthodes de reconstruction portent sur la formulation d'hypothèses sous-jacentes sur la structuration de l'espace des données [8]. Les méthodes de reconstruction comprennent le clustering par k-moyennes [11], les cartes d'auto-organisation (SOM) [8], l'analyse en composantes principales (ACP) [11], les mélanges d'ACP [11], les auto-encodeurs [11].
Dans la littérature, la plupart des méthodes existantes pour la classification monoclasse obtiennent des résultats significatifs [11]. Cependant, ces méthodes montrent quelques faiblesses lorsque peu de données sont disponibles pour l'apprentissage des classificateurs [11]. En particulier, les SVM monoclasse [7], les SVDD [13], les SOM [11], les modèles gaussiens [5], et les modèles de Markov ont besoin de beaucoup d'échantillons de données (environ 40% des données) afin d'affiner plusieurs de leurs paramètres internes [11] pour leur apprentissage.
Compte tenu de notre problème de classification de documents (avec un petit ensemble d'apprentissage), le classificateur monoclasse le plus approprié semble être le modèle d'estimation de densité de Parzen et la méthode k-ppv selon [11]. Nous utiliserons donc un classificateur k-ppv monoclasse qui peut théoriquement travailler à partir de peu de données comme référence dans nos expérimentations. Nous formalisons ce classificateur comme suit :

$$\gamma(x, C) = \begin{cases} 1 & if \quad \sum_{k=1}^{K} \delta(x, t_k) > \frac{K}{2} \\ 0 & sinon \end{cases}$$

$$avec\ \delta(x, t) = \begin{cases} 1 & if \quad \alpha d(x, t) > M \\ 0 & sinon \end{cases}$$

où : M est la distance (euclidienne) moyenne entre tous les exemples de l'ensemble d'apprentissage, $d(x, t)$ est la distance entre x (exemple de la base d'apprentissage) et t (exemple à classer), α est le paramètre de tolérance/rejet et K le nombre de plus proches voisins.
Toutefois, lorsque la variabilité entre les exemples d'une classe est importante, les performances d'un classificateur k-ppv peuvent se dégrader. Une solution alternative pour résoudre cette problématique est d'utiliser les principes de représentation de données symbolique [9]. Par conséquent, le problème de la qualité, de la robustesse et de la fiabilité de cette approximation se pose. L'analyse de données symboliques permet l'analyse non pas d'individus mais de concepts. Un concept est extrait d'une table de données classique, *i.e.* ensemble d'individus décrits par un ensemble d'attributs (qualitatifs et/ou quantitatifs) et représente généralement un sous-ensemble d'individus. Plus précisément, un concept est défini par une "intension" et une "extension". L'intension est un ensemble de propriétés caractéristiques du concept, l'extension est l'ensemble des individus appelés instances du concept qui satisfont ces propriétés [3]. L'ensemble des concepts extraits d'une table de données classique est réuni dans une table appelée "table de données symboliques". Les colonnes d'une telle table sont alors des "attributs symboliques" et les lignes sont appelées "description symbolique" (ou "objet symbolique").
L'analyse de données symbolique a influencé plusieurs domaines. L'un d'eux est la classification où une classe peut être modélisée par un objet symbolique [3]. Dans cet article, nous proposons une approche basée sur la représentation symbolique. Tout d'abord, un ensemble de caractéristiques est extrait à partir de l'ensemble d'apprentissage. Puis, en utilisant les caractéristiques extraites, un vecteur d'intervalles (les valeurs symboliques) est calculé pour représenter la classe cible (ensemble d'apprentissage). Chaque intervalle (donnée symbolique) est calculé en utilisant la moyenne et l'écart-type des valeurs de la caractéristique correspondante. Ces intervalles sont utilisés par les classificateurs monoclasses pour différentes raisons :

- l'intervalle calculé pour chaque caractéristique peut représenter la distribution (moyenne, écart-type) de la classe et prend donc en compte la variabilité intraclasse,
- seulement deux instances sont nécessaires pour pouvoir calculer les moyennes et écart-types nécessaires aux définitions des représentations symboliques,
- une représentation symbolique (*i.e.* un ensemble d'intervalles) peut être calculée pour une nouvelle classe sans engendrer de perturbation sur les autres représentations symboliques.

3 Méthode proposée

Il existe plusieurs représentations de données symboliques [3]. Parmi celles-ci, seule la représentation symbolique par des intervalles peut être calculée à partir d'un faible nombre d'exemples d'apprentissage. Les autres représenta-

tions de données symboliques ont, soit besoin de plusieurs exemples pour être calculées, soit ne sont pas adaptées à nos données car dépendantes d'hypothèses fortes sur la distribution des caractéristiques. Par conséquent, dans ces travaux, nous avons choisi de représenter une classe sous la forme d'un vecteur d'intervalles. Ainsi, chaque vecteur d'intervalles est un classificateur monoclasse.
Pour obtenir l'intervalle associé à une caractéristique, une combinaison de la moyenne statistique et de l'écart-type des valeurs extraites est utilisée [2, 15, 9].
Soit S_j un ensemble de m exemples $(s_{j1}, s_{j2}, \ldots, s_{jm})$ de la classe C_j. F_i représente un vecteur de caractéristiques numériques (de taille n) extrait de l'exemple i de S_j noté s_{ji}

$$F_i = \{f_{i1}, f_{i2}, f_{i3}, \ldots, f_{in}\}$$

Pour chaque caractéristique $f_{.k}$ de chaque classe j, nous calculons la moyenne statistique M_{jk} et l'écart-type σ_{jk}. Les moyennes et écart-types correspondant à chaque caractéristique sont ensuite utilisés pour définir les intervalles F_{jk} selon les formules suivantes :

$$F_{jk} = [F_{jk}^-, F_{jk}^+]$$

$$Avec \begin{cases} F_{jk}^- = M_{jk} - \lambda_j * \sigma_{jk} \\ F_{jk}^+ = M_{jk} + \lambda_j * \sigma_{jk} \end{cases}$$

où F_{jk} représente un intervalle avec pour borne inférieure (F_{jk}^-) et borne supérieure (F_{jk}^+), λ_j un poids dont la valeur est optimisée expérimentalement (*i.e.* λ_j est optimisé par apprentissage, *cf.* équation 1 ci-après).

Le vecteur des n intervalles construits à partir des caractéristiques de la classe C_j correspond alors à la représentation symbolique de cette classe et est noté comme suit :

$$SR_j = \{[F_{j1}^-, F_{j1}^+], [F_{j2}^-, F_{j2}^+], \ldots, [F_{jn}^-, F_{jn}^+]\}$$

Pour classifier un nouvel exemple de test, pour chaque caractéristique, la valeur extraite de l'exemple de test est comparée avec l'intervalle correspondant dans la représentation symbolique de la classe de référence et une mesure de similarité est calculée [2, 15, 9]. Dans notre cas, la similarité $Sim(F_T, SR_j))$ entre l'instance F_T et la référence symbolique SR_j d'une classe particulière j est définie par :

$$Sim(F_T, SR_j) = (\sum_{k=1..n} V_{jk})/n$$

$$V_{jk} = \begin{cases} 1 & if \quad F_{jk}^- \leq f_{tk} \leq F_{jk}^+ \\ 0 & otherwise \end{cases}$$

À partir de la mesure de similarité $Sim(F_T, SR_j)$, un seuil d'acceptation (θ_j) est défini pour chaque classe comme critère de classification. θ_j est calculé automatiquement lors de la phase d'apprentissage afin d'obtenir les meilleurs résultats.
Plus précisément, θ_j et λ_j sont optimisés selon les équations suivantes :

Classe	♯ Images
Attestation_SS (ATT_SS)	60
RIB_PT	16
RIB_GT	76
RIB_MIXT	39
verso Avis impot (IMPOT_VER)	63
Avis impôt (IMPOT_REC)	233
Livret_Famille (LIV_FAM)	36
Images CG anciennes (CG)	21

TABLE 1 – Répartition du nombre d'images dans chacune des classes

$$\theta_j^* = majorityvote(\theta_j) \quad (1)$$

$$\lambda_j^* = mean(\lambda_j \in \Lambda) \quad (2)$$

Avec Λ l'ensemble des λ_j pour lesquels $\theta_j = \theta_j^*$ et le nombre de documents acceptés à tort est nul (*i.e.* $F_p = 0$). La valeur optimale de (θ_j) et la valeur moyenne de (λ_j) correspondante sont sélectionnées avec pour objectif de minimiser le nombre de faux positifs, *i.e.* avec un taux élevé de rejet afin d'obtenir la plus faible probabilité de retrouver des documents négatifs (confusion).

4 Expérimentations et Analyse comparative

4.1 Données et métriques utilisées

Pour évaluer notre méthode, nous avons choisi de l'appliquer à une tâche de classification d'images de documents. Nous avons mené plusieurs expérimentations sur une base composée de 544 images de documents issues d'un problème réel. Ces données ont été catégorisées en 8 classes par les experts de notre partenaire industriel. La Figure 1 illustre deux exemples d'images issues de deux classes différentes. Le Tableau 1 résume les intitulés des classes ainsi que le nombre de documents qu'elles contiennent.
Pour évaluer la performance de notre système, nous utilisons les critères de rappel (*Rec.*) et de précision (*Pre.*). Nous rappelons que ces mesures sont calculées comme suit :

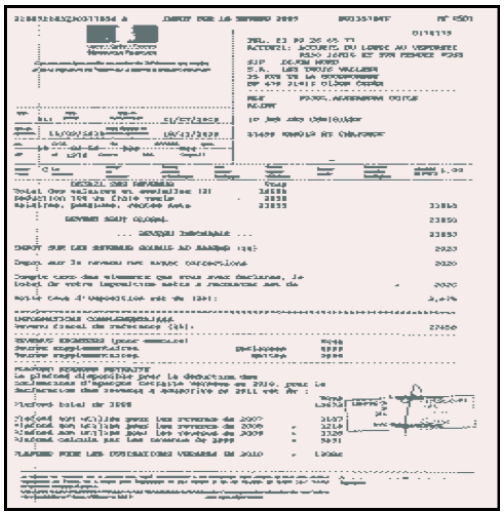
$$Pre. = T_p/(T_p + F_p)$$

$$Rec. = T_p/(T_p + F_n)$$

où T_p est l'ensemble des documents correctement classés et T_n l'ensemble des documents rejetés à raison. F_n désigne l'ensemble des documents qui appartiennent à la classe et ont été rejetés, F_p les documents acceptés à tort.

4.2 Résultats expérimentaux

Pour améliorer la qualité des images de documents avant l'extraction de caractéristiques, un certain nombre de techniques de prétraitement (recadrage, réalignement, débruitage, et rotation automatique) ont été appliquées sur les



(a) IMPOT_REC

(b) RIB_PT

FIGURE 1 – Exemples de documents issus de deux classes différentes

images de documents. Pour caractériser chaque image de document, un vecteur de 101 caractéristiques a été calculé pour chaque image du document. Ces caractéristiques sont celles utilisées classiquement, comme le rapport entre la hauteur et la largeur, la densité des niveaux de gris, les histogrammes de couleurs, l'histogramme des projections horizontales et verticales, le nombre de composantes connexes, le nombre d'occlusions, l'histogramme des longueurs des lignes droites, des histogrammes sur la taille des composantes connexes et les histogrammes sur les orientations des contours. Toutes ces caractéristiques ont été calculées sur les pages entières, mais aussi localement dans différentes régions des images (*cf.* [10]).
Comme la base de données utilisée pour l'expérimentation est composée de huit classes, huit classificateurs symboliques monoclasses indépendants ont été construits et utilisés. Pour l'apprentissage du classificateur associé à une classe, $m = 10$ images de classe ont été choisies aléatoirement pour former l'ensemble d'apprentissage. Le reste des documents dans chaque classe est considéré comme les documents appartenant à l'ensemble test, ce qui représente un ensemble de 464 documents. Pour représenter chaque classe avec un objet symbolique, nous calculons l'ensemble d'intervalles de caractéristiques en utilisant $n = 101$ caractéristiques.
Les résultats obtenus en utilisant le classificateur monoclasse proposé basé sur la représentation symbolique sont présentés dans le tableau 2. Nous rappelons que, dans l'expérimentation, les seuils (θ_j) et (λ_j) sont choisis expérimentalement selon les équations 1 et 2.

	λ	θ	T_p	T_n	F_p	F_n	Pre.	Rec.
ATT_SS	2,0	0,91	23	414	0	27	100	46
RIB_PT	1,4	0,80	6	458	0	0	100	100
RIB_GT	2,2	0,96	16	398	0	50	100	24,2
RIB_MIXT	2,2	0,95	9	435	0	20	100	31
IMPOT_VER	1,7	0,98	16	411	0	37	100	30,1
IMPOT_REC	1,8	0,95	91	241	0	132	100	40,8
LIV_FAM	1,7	0,96	14	438	0	12	100	53,8
CG	1,1	0,86	4	453	0	7	100	36,3

TABLE 2 – Résultats obtenus par le classificateur à base de représentation symbolique avec un vecteur de 101 caractéristiques.

La sélection des 10 exemples pour former l'ensemble d'apprentissage de chaque classe peut avoir une influence sur les résultats. Par conséquent, nous avons décidé de confirmer nos résultats par un processus de validation croisée. Les résultats obtenus par le classificateur monoclasse proposé dans ce contexte sont présentés dans le tableau 3. Pour chaque expérimentation, 10 documents ont été choisis aléatoirement comme exemples d'apprentissage et les données restantes ont été considérées comme l'ensemble de test. Cette expérimentation a été réalisée 10 fois et le résultat final pour chaque classe correspond à la moyenne des 10 résultats obtenus.
Dans le tableau 3, la moyenne des vrais positifs (T_p) est de 48,10%, ce qui représente une amélioration d'environ 10% par rapport aux résultats obtenus précédemment (38,5%, cf. Tableau 2). En contrepartie, la moyenne de faux positifs a atteint 8,5% (0% précédemment).
L'un des processus existant pour améliorer la qualité de la classification est la sélection des caractéristiques importantes et la suppression des non pertinentes en amont de la construction du classificateur. Bien que la sélection de caractéristiques dans les problèmes de classification multiclasse ait fait l'objet de beaucoup de recherches, peu de méthodes de réduction de dimension sont disponibles pour une utilisation dans les problèmes de classification monoclasse ([4]). En effet, l'évaluation du pouvoir discriminant des caractéristiques est un problème difficile quand aucun exemple négatif n'est disponible. Partant de ce constat, nous vous proposons d'estimer la stabilité (robustesse) des caractéristiques par le calcul d'un score basé sur deux indices :

	T_p	T_n	F_p	F_n	Pre.	Rec.
ATT_SS	26,1	411,4	2,6	23,9	90,94	52,20
RIB_PT	5,9	456,8	1,2	0,1	83,10	98,33
RIB_GT	31,3	393	5	34,7	86,23	47,42
RIB_MIXT	11	422,5	12,5	18	46,81	37,93
IMPOT_VER	26,4	403,7	7,3	26,6	78,34	49,81
IMPOT_REC	104	238,8	2,2	119	97,93	46,64
LIV_FAM	12,1	433,9	4,6	13,9	72,46	46,54
CG	6,6	448,9	4,1	4,4	61,68	60,00

TABLE 3 – Résultats obtenus par le classificateur à base de représentation symbolique en suivant un processus de validation croisée avec un vecteur de 101 caractéristiques.

- P : pourcentage des valeurs dont la dispersion autour de la moyenne est d'un écart type.
- R : rapport entre l'écart inter-quartile (différence entre le troisième et le première quartile) et l'étendue des données (différence entre les valeurs maximales et minimales)

A partir de ces deux valeurs, un score est calculé tel que $S = P + (1 - R)$. Ce score varie entre 0 et 2 (2 signifiant que les valeurs de cette caractéristique sont identiques pour l'ensemble des données). Ainsi, un classement des caractéristiques en fonction de leur stabilité pour une classe est possible. Cette méthode permettra une réduction de l'ensemble des caractéristiques en sélectionnant les N les plus robustes. Ce critère de sélection, qui tend à sélectionner les caractéristiques ayant une distribution normale sur une classe, est particulièrement adapté à notre modèle symbolique pour lequel les intervalles sont construits sur la base de la moyenne et de l'écart-type.
Nous avons reproduit les expérimentations précédentes en utilisant un vecteur de caractéristiques réduit. Les résultats sont reportés respectivement dans les tableaux 4 et 5. Pour chacune des expérimentations, le nombre de caractéristiques retenus (n) est optimisé expérimentalement.

	n	λ	θ	T_p	T_n	F_p	F_n	Pre.	Rec.
ATT_SS	60	2,1	0,90	25	414	0	25	100	50
RIB_PT	30	2,1	0,87	6	458	0	0	100	100
RIB_GT	30	2,3	0,94	44	398	0	22	100	66,6
RIB_MIXT	40	1,9	0,88	10	435	0	19	100	34,4
IMPOT_VER	65	1,7	0,96	19	411	0	34	100	35,8
IMPOT_REC	50	1,4	0,85	105	241	0	118	100	47
LIV_FAM	40	1,7	0,93	18	438	0	8	100	69,2
CG	80	1,1	0,84	6	453	0	5	100	54,5

TABLE 4 – Résultats obtenus par le classificateur à base de représentation symbolique avec un vecteur de caractéristiques réduit.

Nous pouvons remarquer que la réduction du vecteur de caractéristiques entraîne une amélioration notable des résultats. D'après le tableau 4, il apparaît que, en moyenne, 50,2% des documents sont classés correctement lorsque le pourcentage de précision dans toutes les classes reste à 100% (aucun faux positif). Les résultats de l'expérimentation par une validation croisée sont présentés dans le tableau 5. D'après ce tableau, la moyenne des vrais positifs est de 56,10%, soit une amélioration de 6% par rapport aux résultats obtenus précédemment. Cette différence s'explique par la sélection aléatoire des 10 exemples constituant l'ensemble d'apprentissage (lors de l'expérimentation précédente), ce qui peut induire une mauvaise représentation de la classe.

	n	T_p	T_n	F_p	F_n	Pre.	Rec.
ATT_SS	70	26,8	411,8	2,2	23,2	92,41	53,60
RIB_PT	30	5,8	457	1	0,2	85,29	96,67
RIB_GT	30	44,4	397,8	0,2	21,6	99,55	67,27
RIB_MIXT	70	11,8	421,6	17,8	17,2	39,86	40,69
IMPOT_VER	70	25,4	404	7	27,6	78,40	47,92
IMPOT_REC	80	129,4	237,4	3,6	93,6	97,29	58,03
LIV_FAM	60	11,4	433,4	4,6	14,6	71,25	43,85
CG	40	5,4	450,4	2,6	5,6	67,50	49,09

TABLE 5 – Résultats obtenus par le classificateur à base de représentation symbolique en suivant un processus de validation croisée avec un vecteur de caractéristiques réduit.

4.3 Analyse Comparative

Pour avoir une analyse comparative des résultats avec une autre méthode de classification monoclasse, nous avons utilisé les résultats (présentés dans le Tableau 6) d'un k-ppv monoclasse (cf. section 3) appliqué sur le même ensemble de données. 10 exemples de chaque catégorie ont été choisis pour l'apprentissage du k-ppv monoclasse. De plus, les paramètres K, N (taille du vecteur) et α sont optimisés. Le paramètre α correspond au coefficient du seuil de distance limite, obtenu à partir des distances moyennes entre les exemples d'apprentissage.
Pour le classificateur k-ppv monoclasse, il est à noter que la sélection de caractéristiques pour chaque classe fournit de meilleurs résultats de classification. A partir des résultats présentés dans le tableau 6, nous avons calculé une moyenne des précisions égale à 28,8%.
La moyenne des vrais positifs obtenus dans les différentes expérimentations est présentée dans le tableau 7. D'après le tableau 7, nous pouvons noter que la technique proposée fournit de meilleurs résultats (amélioration de 28%) par rapport à la méthode de classification k-ppv monoclasse.
La méthode proposée a également été appliquée à d'autres bases de données (publiques) et les résultats comparatifs avec différentes méthodes [14, 1] sont affichés dans le Tableau 8. *Iris (versicolor)* contient 50 objets positifs et 100 objets négatifs, décrits par un vecteur de 4 caractéristiques. Etant donné ce faible nombre, le résultat de notre méthode est légèrement inférieur aux résultats des autres méthodes. *Breast Wisconsin Dataset (malignant)* contient 241 objets

	n	α	T_p	T_n	F_p	F_n
ATT_SS	50	1,2	18	414	0	32
RIB_PT	30	0,9	6	458	0	0
RIB_GT	60	0,8	17	397	0	49
RIB_MIXT	70	0,8	10	425	0	19
IMPOT_VER	70	0,9	22	411	0	31
IMPOT_REC	60	1,0	54	251	0	169
LIV_FAM	50	0,7	3	438	0	23
CG	60	0,7	4	453	0	7

TABLE 6 – Résultats obtenus par le classificateur k-ppv avec un vecteur de caractéristiques réduit.

Methode	Vecteur carac.	Apprentissage	Moyenne (%)
k-ppv	Vecteur réduit	10 instances	28,8
Méthode proposée	Vecteur complet	10 instances	38,5
		val. croisée	48,1
	Vecteur réduit	10 instances	50,2
		val. croisée	56,1

TABLE 7 – Moyenne des vrais positifs obtenus dans les différentes expérimentations.

positifs et 458 objets négatifs, décrits par un vecteur de 9 caractéristiques. En utilisant la méthode proposée, nous avons obtenu des résultats supérieurs par rapport à ceux des autres méthodes [1], comme indiqué dans le Tableau 8.

	Gaussian	Parzen	k-moyennes	k-ppv	Méthode proposée
Iris	99,4	99,0	98,4	98,4	97,2
Breast	82,3	72,3	84,6	69,3	98,87

TABLE 8 – Comparaison de différentes méthodes sur les bases de données *Iris* et *Breast*.

5 Conclusion et Perspectives

Cet article présente un nouveau classificateur monoclasse basé sur une représentation symbolique. La méthode s'appuie sur la construction d'intervalles en utilisant la moyenne et l'écart-type des valeurs du vecteur de caractéristiques. Comme cette méthode est appropriée pour un faible nombre d'exemples d'apprentissage, elle peut être utilisée efficacement pour résoudre le problème de classification d'images de documents dans un contexte industriel. La technique proposée a une complexité très faible et atteint de bonnes performances en précision par rapport aux autres méthodes de classification monoclasse existantes.
Dans de futurs travaux, nous souhaiterions réaliser une étude sur le nombre optimal d'exemples d'apprentissage nécessaire pour l'obtention de résultats satisfaisants. En outre, des connaissances *a priori* sur les classes pourraient également être intégrées à l'ensemble des caractéristiques pour améliorer les performances de classification. De plus, une méthode de sélection des exemples d'apprentissage les plus pertinents est à l'étude car, comme le montrent les tables 2 et 3, la méthode reste sensible aux exemples d'apprentissage utilisés. Enfin, un renforcement de la représentation symbolique est envisagé par construction de plusieurs vecteurs d'intervalles pour une même classe, avec pour objectif d'affiner la représentation et de limiter la confusion entre les classes. Cette approche a pour but de mieux prendre en compte les classes de documents hétérogènes.

Références

[1] http ://homepage.tudelft.nl/n9d04/occ/index.html.

[2] A. Alaei, P. Nagabhushan, and U. Pal. Persian/Arabic handwritten numeral recognition : An approach based on symbolic representation. In *International Conference on Signal and Image Processing*, pages 435–439, 2009.

[3] E. Diday and F. Esposito. An introduction to symbolic data analysis and the sodas software. *Intelligent Data Analysis*, 7(6) :583–601, 2003.

[4] I. Fodor. A survey of dimension reduction techniques. Technical report, 2002.

[5] M. Kemmler, E. Rodner, and J. Denzler. One-class classification with gaussian processes. In Ron Kimmel, Reinhard Klette, and Akihiro Sugimoto, editors, *Computer Vision – ACCV 2010*, volume 6493 of *Lecture Notes in Computer Science*, pages 489–500. Springer Berlin Heidelberg, 2011.

[6] S.S. Khan and M.G. Madden. A survey of recent trends in one class classification. In *Artificial Intelligence and Cognitive Science*, Lecture Notes in Computer Science, pages 188–197. Springer Berlin Heidelberg, 2010.

[7] L.M. Manevitz and M. Yousef. One-class svms for document classification. *J. Mach. Learn. Res.*, 2002.

[8] O. Mazhelis. One-class classifiers : a review and analysis of suitability in the context of mobile-masquerader detection. *South African Computer Journal*, 2006.

[9] H.N. Prakash and D.S. Guru. Offline signature verification : An approach based on score level fusion. *International Journal of Computer Applications*, 1(1) :52–58, 2010.

[10] N. Sidère, J.-Y. Ramel, S. Barrat, V. Poulain D'Andecy, and S. Kebairi. Identification de documents par classification monoclasse. In *Actes du treizième Colloque International Francophone sur l'Ecrit et le Document ; CIFED 2014*, page 277–290, 2014.

[11] D.M.J. Tax. *One-class classification : Concept learning in the absence of counter-examples*. PhD thesis, Technische Universiteit Delft, 2001.

[12] D.M.J. Tax and R.P.W. Duin. Data domain description using support vectors. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 251–256, 1999.

[13] D.M.J. Tax, R.P.W. Duin, N. Cristianini, J. Shawe-Taylor, and B. Williamson. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, 2 :155–173, 2001.

[14] F. Van Der Heijden, R.P.W. Duin, D. de Ridder, and D.M.J. Tax. *Classification, parameter estimation and state estimation - an engineering approach using Matlab*. John Wiley & Sons, 2004.

[15] T.N. Vikram, K.C. Gowda, and S.R. Urs. Symbolic representation of kannada characters for recognition. In *Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference on*, pages 823–826, 2008.